CMS Conditions Data Access using FroNTier

Lee Lueking
CMS Offline Software and Computing

5 September 2007

CHEP 2007 – Distributed Data Analysis and Information Management

Outline

- Motivation
- Implementation Details
- Deployment Overview
- Performance

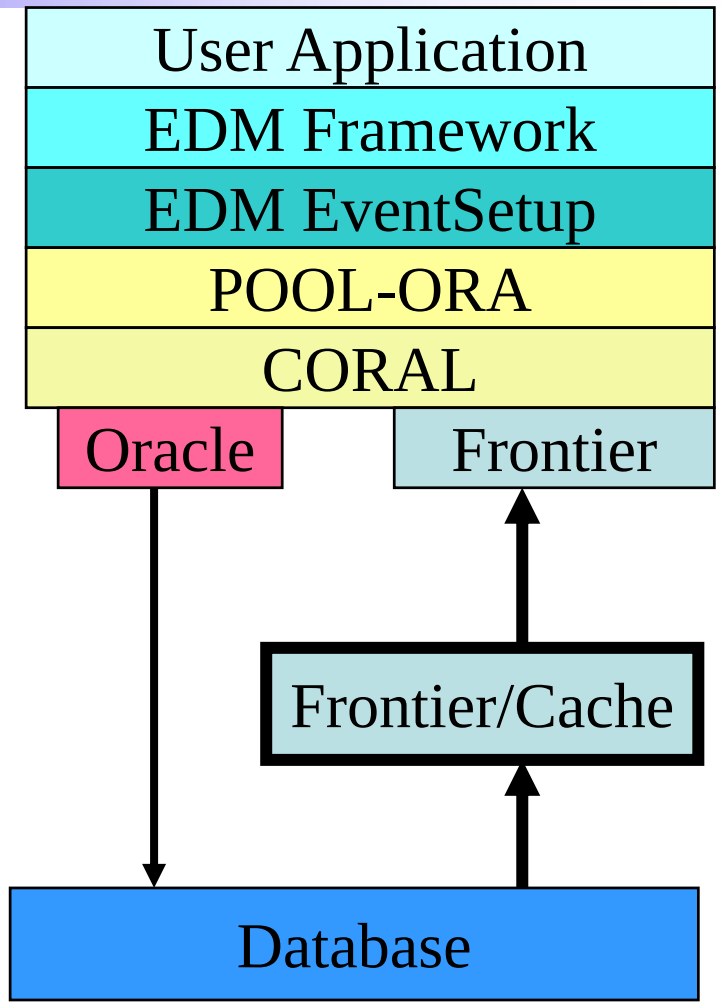
Acknowledgements
•The Frontier Team: Barry Blumenfeld (JHU), David Dykstra (FNAL), Eric Wicklund (FNAL)

Motivation

- CMS ***conditions data*** includes calibration, alignment, and configuration information used for offline detector ***event data*** processing.
- Conditions data is keyed by time (run number) and defined to be immutable, i.e. new entries require new tags (versions).
- A given object may be used by thousands of jobs. Caching such info close to the processing activity provides significant performance gains.
- Readily deployable, highly reliable and easily maintainable web proxy/caching servers are a logical solution.

CMS Software Stack

- POOL-ORA (Object Relational Access) is used to map C++ objects to Relational schema.
- A CORAL-Frontier plugin provides read-only access to the POOL DB objects via Frontier.



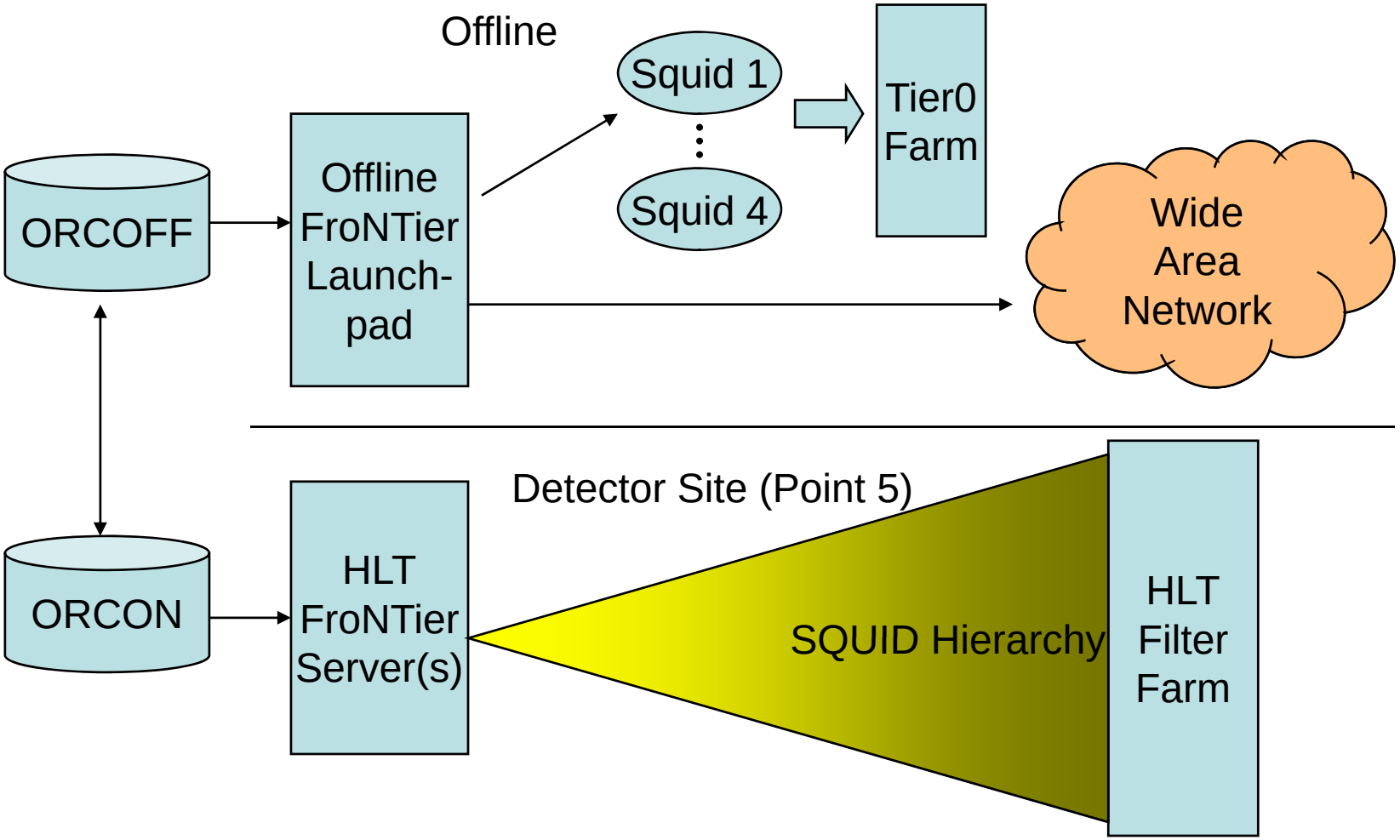
Implementation

- Pool and CORAL generate SQL queries from the CMSSW C++ objects.
- The FroNTier client converts the SQL into an HTTP GET and sends it over the network to the FroNTier server.
- The FroNTier server, a servlet under Tomcat, unpacks the SQL request, sends it to the DB server, and retrieves the needed data.
- The data is optionally compressed, and then packed into an HTTP formatted stream sent back to the client.
- Squid proxy/caching server(s) between the FroNTier server and client caches requested objects, significantly improving performance and reducing load on the DB.

Advantages

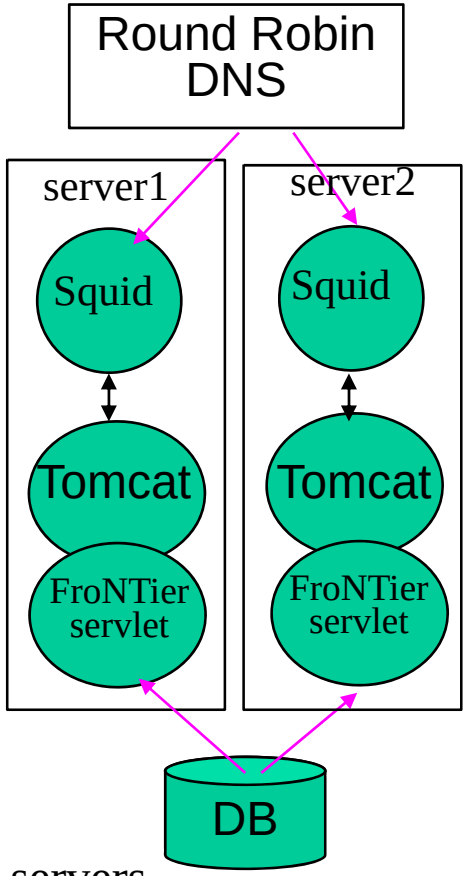
- The system uses standard tools
  - Highly reliable – Tomcat, Squids well proven
  - Easy installation – Tar ball and script
  - Highly configurable – Customize to site environment, security, etc.
  - Well documented – Books and web sites
  - Readily monitored – SNMP, MRTG, awstats
- Easy administration at Tier-N centers
  - No DBA's needed beyond central DB @CERN
  - Caches are loaded on demand and self managing

Overview


Offline
Squid 1
Squid 4
Tier0 Farm
ORCOFF
Offline FroNTier Launch-pad
Wide Area Network
Detector Site (Point 5)
ORCON
HLT FroNTier Server(s)
SQUID Hierarchy
HLT Filter Farm

FroNTier “Launchpad”

- Squid caching proxy
  - Load shared with Round-Robin DNS
  - Configured in “accelerator mode”
  - “Wide open frontier”*
  - Peer caching removed because it was incompatible with collapsed forwarding.
- Tomcat - standard
- FroNTier servlet
  - Distributed as WAR (Web ARchive)
    - Unpack in Tomcat webapps dir
    - Change 2 files if name is different
  - One xml file describes DB connection



* The squids in the launchpad ONLY talk to the Frontier Tomcat servers. No registration or ACL’s required.

Cache Coherency (1/2)

Metadata

Name (tag)	POOL Token
“Test”	To container A
“Online”	To container B
“prod”	To container C
“stuff”	To Container D

Lower end of run range

Container B

Conditions IOV

Run	POOL Token
100	To payload 1
200	To payload 2
300	To payload 3
500	To payload 4
New Run	New payload Appended

- So-called “metadata” for conditions data. These are names or “tags” that refer to a specific set of IOV’s (Interval Of Validity) and associated payloads in the pool-ora repository.
- By decree, data in Conditions IOV can NOT change.
- Therefore, caching is OK.

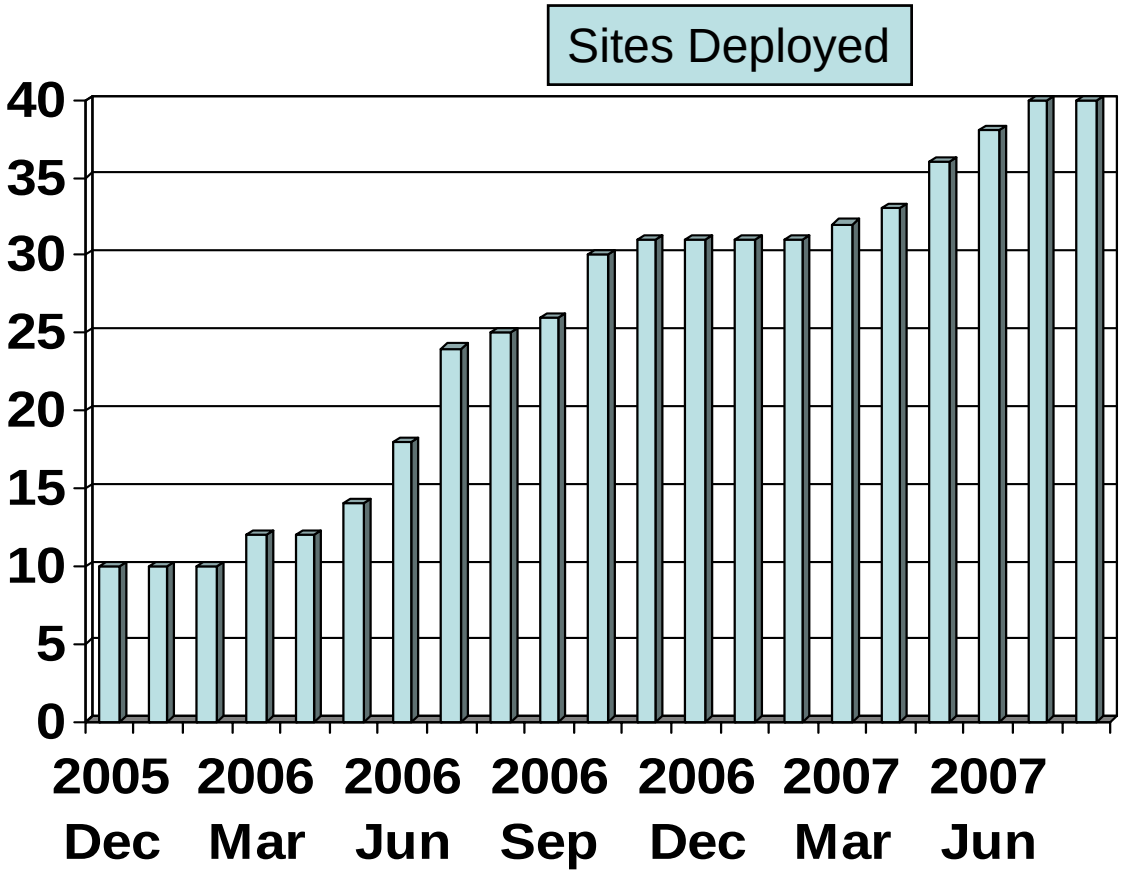
- BUT…
- New IOV’s and payloads can be appended to in order to extend the IOV range.
- This is NOT OK for caching.

Cache Coherency (2/2)

- After considering many ideas, adopted following solution:
  - All objects that are cached have expiration times
  - Shortlived: Metadata objects, including the pointers to payload objects, expire on a short time period
  - Longlived: Payload objects have a long expiration time.
- The values of the short and long times are adjusted according to where the data is being used. For example:
  - Online: the calibrations change quickly as new data is added for upcoming runs.
  - Tier 0: calibrations change on the order of a few hours as new runs appear for reconstruction.
  - Tier 1 +: Conditions data may be stable for weeks.
- The value of the short and long expirations is modified at the central FroNTier server, so they can easily be tuned as needed.

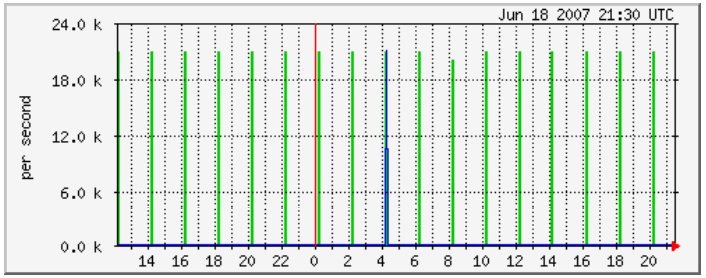
Squid Deployment Status

- Late 2005, 10 centers used for testing
- Additional installation May through Oct. 2006 used for CSA06
- Additional 20-30 sites for CSA07 possible
- Very few problems with the installation procedures CMS provides.



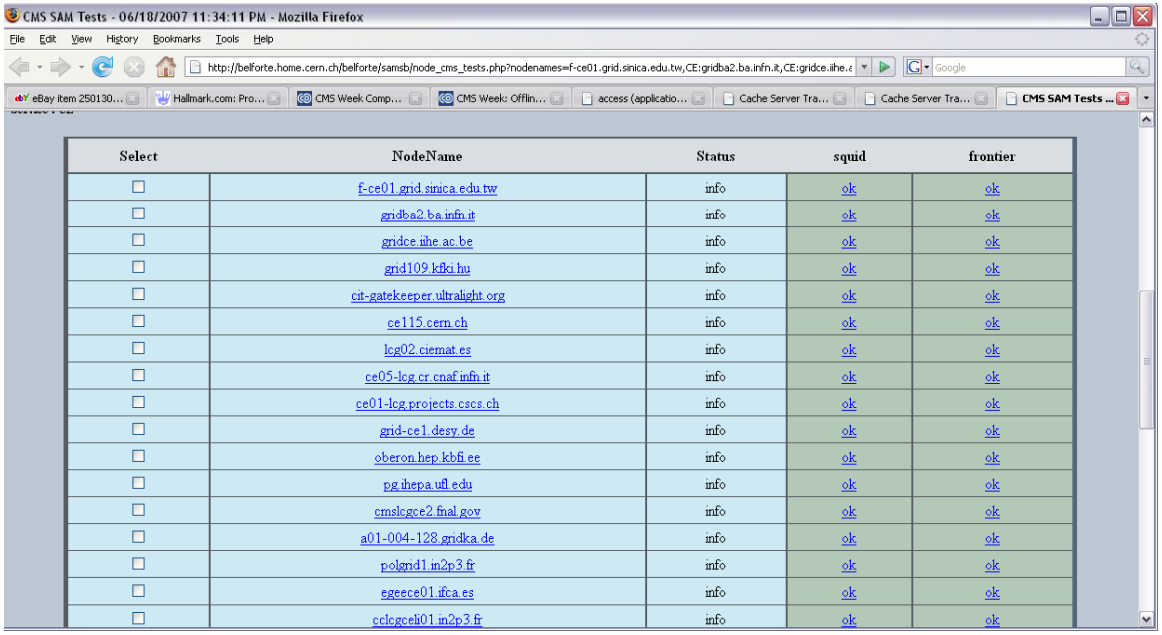
Service Availability Monitoring (SAM)

Eric Wicklund/ Stefano Belforte



- Heartbeat monitoring tests are run every 2 hours
- 40 sites included
- http://www-cdf.fnal.gov/~wicklund/squid.html

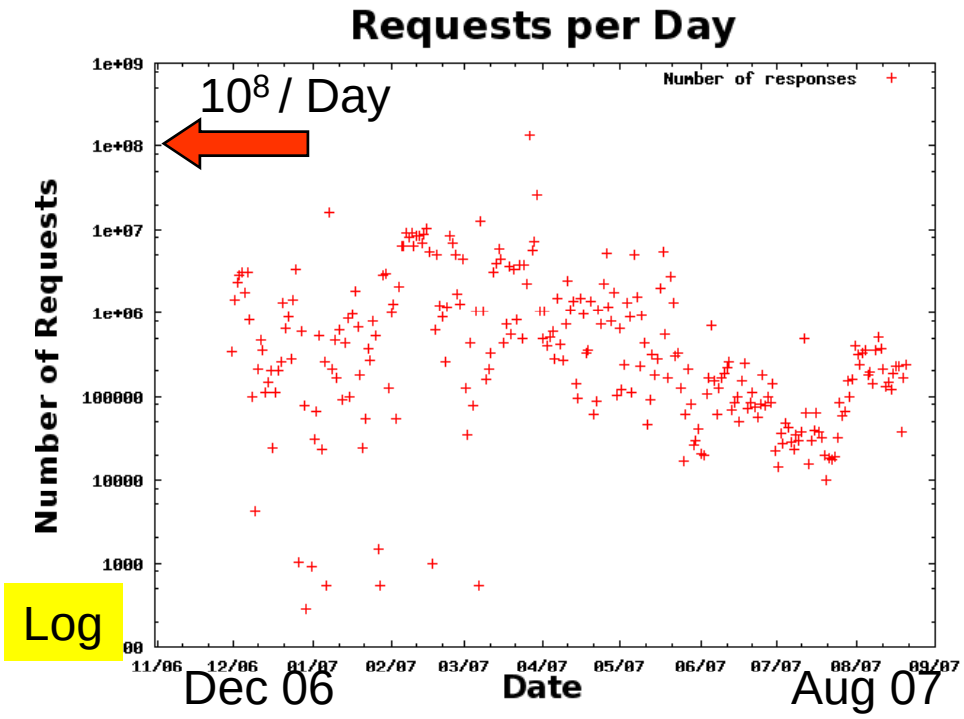
- Check integrity of local squid, and
- Overall Frontier system and connection to CERN

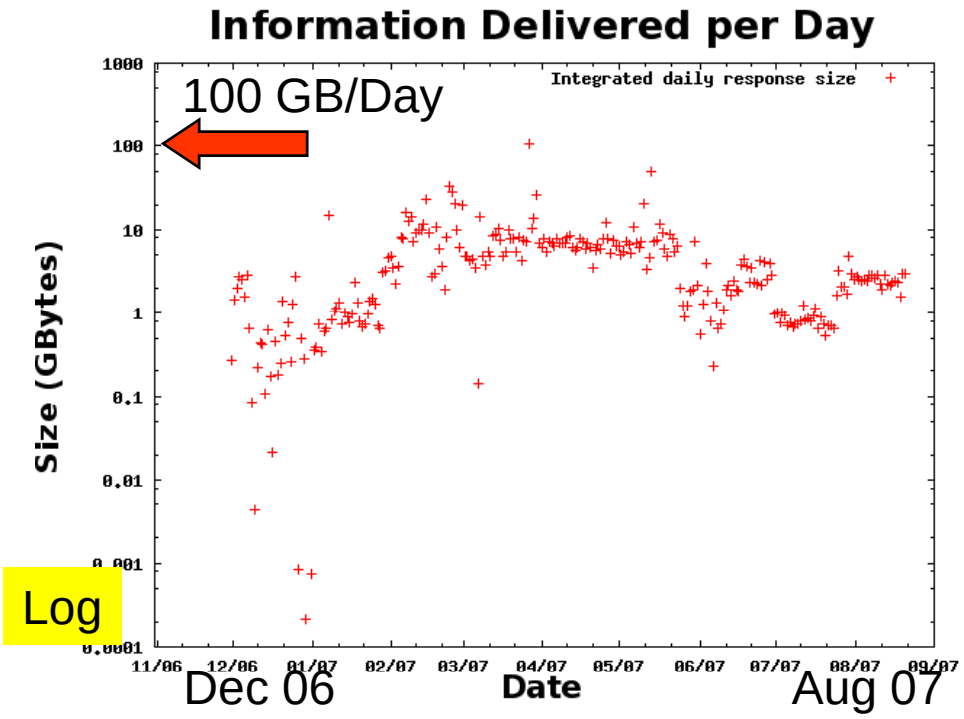


Select	NodeName	Status	squid	frontier
☐	f-ce01.grid.sinica.edu.tw	info	ok	ok
☐	gridba2.ba.infn.it	info	ok	ok
☐	gridce.iihe.ac.be	info	ok	ok
☐	grid109.kfki.hu	info	ok	ok
☐	cit-gatekeeper.ultralight.org	info	ok	ok
☐	ce115.cern.ch	info	ok	ok
☐	lcg02.ciemat.es	info	ok	ok
☐	ce05-lcg.cr.cnaf.infn.it	info	ok	ok
☐	ce01-lcg.projects.cscs.ch	info	ok	ok
☐	grid-ce1.desy.de	info	ok	ok
☐	oberon.hep.kbfi.ee	info	ok	ok
☐	pg.ihepa.ufl.edu	info	ok	ok
☐	cmslcgce2.fnal.gov	info	ok	ok
☐	a01-004-128.gridka.de	info	ok	ok
☐	polgrid1.in2p3.fr	info	ok	ok
☐	egeece01.ifca.es	info	ok	ok
☐	cclcgceli01.in2p3.fr	info	ok	ok

Launchpad Operation

All activity on Launchpad: Production, Development and Testing





- Number of daily objects varies widely
- Peak day (10**8) was fail-over from local squid at Bari, good test of infrastructure.

- Object size depends on type of activity occurring

Specific Challenges

- HLT (High Level Trigger)
  - Startup time for Cal/Ali < 10 seconds.
  - Simultaneous
  - Uses hierarchy of squid caches
- Tier0 (Prompt Reconstruct)
  - Startup time for conditions load < 1% of total job time.
  - Usually staggered
  - DNS Round Robin should scale to 8 squids

Parameter	HLT	Tier0
# Nodes	2000	1000
# Processes	~16k	~3k
Startup	<10 sec all clients	<100 sec per client
Client Access	Simultaneous	Staggered
Cache Load	< 1 Min	N/A
Tot Obj Size	100 MB*	150 MB*
New Objects	100% / run*	100% / run*
# Squids	1 per node	Scalable (2-8)

* Worst case scenario

Starting Many Jobs Simultaneously

Online HLT Problem

- All nodes start same application at the same time
- Pre-loading data must be < 1 minute
- Loading data to jobs must be < 10 seconds
- Estimating 100MB of data, 2000 nodes, 8 jobs/node
  - 100 * 2000 * 8 = 1.6TB
- Asymmetrical network
  - Nodes organized in 50 racks of 40 nodes each
  - non-blocking gigabit intra-rack, gigabit inter-rack

Starting Many Jobs Simultaneously

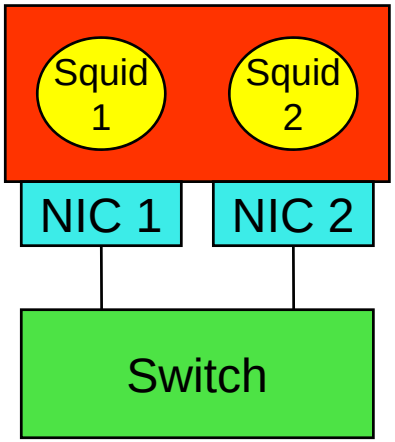
Online HLT Solution

- Configured to pre-load simultaneously in tiers
- Each squid feeding 4 means 6 tiers for 2000 nodes
  - 50 racks reached in 3 tiers, 3 tiers inside each rack
- Measurements on test cluster indicate requirements can be met
  - bottleneck becomes the conversion from DB to httpin FroNTier server
  - 10-second loading always reads from pre-filled local squid

Testing w/ Multiple NICs

Potential Tier-0 and Tier-1 Augmentation

- On multi-processor/multi-core machines CPU resources under utilized. Using multiple network interfaces can resolve this.
- Two approaches tried
  - Multi-homed (2 ip addresses): machine looks like multiple nodes
  - Bonded interfaces (1 ip address): multiple NICs used together to increase throughput.
- Using Bonded approach at FNAL
  - Works best w/ specific load balancing approach
  - Able to improve throughput by factor of two over single unbonded NIC
  - Requires 2 squids on same server machine, ~200 MBps (Squid single threaded).
- Many sites, for now, prefer just adding more machines to make network management simpler. This may change.



Summary

- FroNTier is used by CMS for all access to all conditions data.
- The ease of deployment, stability of operation, and high performance make the FroNTier approach well suited to the GRID environment being used for CMS offline, as well as for the online environment used by the CMS High Level Trigger (HLT).
- The use of standard software, such as Squid and various monitoring tools, make the system reliable, highly configurable and easy to maintain.
- We have gained significant operational experience over the last year in CMS, there are currently 40 squid sites being monitored and many more expected.

Finish

Example Calibration and Alignment Object Sizes (Monte Carlo)

- Table shows examples of the size of conditions data for a few sub-detectors
- Zipping of data can significantly reduce the size of network transfers, at the cost of some server and client performance.
  - Online and transfers local to CERN it does not help
  - For sites remote to CERN it is useful
- Object sizes and zipping factors for real detector data may be quite different than for MC

Detector Sub-System (not all systems included)	Data size (MB)	
	Non-compressed	Compressed (zipped)
HCAL	1	0.4
ECAL	7	3.2
Drift Tubes	12	?
Si Track	20	?
Pixel Track	130	?
Current Total in MC	280	?

Compression factors unrealistic for MC data